様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　５　月　３１　日現在

機関番号：　１２６０１
研究種目：　基盤研究(B)（一般）
研究期間：　2013～2015
課題番号：　２５２８０１１１

研究課題名（和文）マルチソーシャルメディアにおける情報伝播挙動の類型化に関する研究

研究課題名（英文）Classifying Behaviors of Information Cascades on Multi Social Media

研究代表者
豊田　正史（TOYODA, Masashi）

東京大学・生産技術研究所・准教授

研究者番号：６０４４７３４９

交付決定額（研究期間全体）：（直接経費）　　14,000,000円

研究成果の概要（和文）：ソーシャルメディアの重要な特徴の一つは友人間のネットワークを通して情報が連鎖的に伝播していくことにある。本研究では、多様な話題に関する情報伝播を分析し、コンテンツ、伝播経路、参加ユーザ、使用言語など伝播挙動に影響を与える特徴量を詳細に調査した。分析結果を用い、社会的影響力に基づく話題の分類手法や多言語間での情報伝播の予測手法を提案し、情報伝播挙動を可視化するシステムを開発した。

研究成果の概要（英文）：Information cascade through friendship networks is one of the most important phenomena on social media. We analyzed information cascades on heterogeneous topics, and investigated how their cascading behaviors affected by features, such as contents, paths of cascades, involved users, and languages. Based on the analysis, we proposed methods for classifying cascades by their social influence, and for predicting occurrence of cross-lingual cascades. Finally, we developed a system for visualizing cascading behaviors.

研究分野：　情報学

キーワード：　ソーシャルウェブ　社会ネットワーク分析

１．研究開始当初の背景

World Wide Web (以下 Web と略記) におけるソーシャルメディアのユーザは、米国 comScore 社の調査によると 2011 年に 12 億人を超え、人々のオンラインでの活動時間のうち 20%を費やすまでに至っている[1]。特に、ユーザが友人との情報共有を円滑に行えるようにするソーシャルネットワークサービスや、友人間で短いメッセージを公開し共有できるマイクロブログサービスなどの隆盛が顕著である。ソーシャルメディアの利用時間は既に検索エンジンや電子メールサービスの利用時間を上回っており、日常的に利用される主要な情報源の一つとなっている。

ソーシャルメディアの大きな特徴の一つは友人間のネットワークを通して情報が伝播することにある。日々多様な話題がソーシャルメディア上を流れているが、その伝播の挙動は話題の性質によって異なる。例えば、昨年の東日本大震災の直後にマイクロブログを起点とする節電の呼びかけが起こったが、事態の緊急性もあり非常に早いスピードで多くの人々に広まった。一方、原発事故に関する話題は、事態が長期化するにつれ強い関心を持つ比較的少数の人々の間で議論が継続的に行われている。このような話題の性質による情報伝播の挙動の違いを把握することは、災害や事故に対する人々の反応を観測することのみならず、政府、自治体、企業等の広報活動やキャンペーンの効果を調査する上で重要である。

ソーシャルメディア上の話題と情報伝播挙動の関連については、話題に関する言及数の時系列パターンを分析した研究や、ソーシャルネットワークにおける情報伝播の経路及びタイミングの側面を分析した研究などが近年進められている。しかし、これら様々な情報伝播の特徴が各種話題の性質とどのように関連しているのか、また、メディア間にまたがる情報伝播の挙動や、メディアによる挙動の違いについては、明らかになっていなかった。

２．研究の目的

本研究は、複数のソーシャルメディア上における情報伝播の挙動を類型化することのできる特徴量の組合せを抽出し、これを用いて話題の性質を表す情報伝播挙動のプロファイルを作成する方法を確立することを目的とし、以下のテーマを実施する。

（１） 個々のソーシャルメディアにおいて、話題に参加している各ユーザの影響力、影響され易さ、発言回数、ユーザ間ネットワークの形状、および情報の伝播経路など様々な特徴量を話題毎に分析する。

（２）ブログやマイクロブログなど複数のソーシャルメディア間における情報伝播を考慮した特徴量の拡張を行い、メディア間の話題出現の時間差、情報伝播経路などが上記パターンに影響するかどうかを明らかにする。また、特徴量の時間的な変化を分析し、話題の質的な変化を抽出可能であるかどうかを明らかにする。

（３）上記特徴量を総合してソーシャルメディア上の話題の性質を説明可能な情報伝播のプロファイルを構成する。アーカイブ上の主要な話題に関して情報伝播プロファイルの検索・閲覧システムを構築し、過去の類似事例を適切に可視化するとともにインタラクティブに探索可能とする手法を構築する。

３．研究の方法

分析対象としては、これまでに収集してきたブログデータ、およびマイクロブログ (Twitter.com) のデータを用いる予定であった。ブログデータについては 2006 年より数百万のブログから数十億記事規模、Twitter データについては 2011 年より数百万ユーザのつぶやきを数百億収集しており、多様な分野の話題を分析するのに十分なデータを整えた。しかし、研究開始時にはブログ空間における情報伝播の影響力が著しく弱くなっており、ブログ記事は Twitter や Facebook 等他のメディアへリンクをポストすることにより拡散するケースが多くなっていた。このため、本研究においては、大規模収集を行っている Twitter における情報伝播の多様性、および言語コミュニティにまたがる情報伝播に着目して研究を進めることとした。

情報伝播の多様性に関しては、商品や人物に対する影響の大きい批判や、事故災害情報の周知等、情報伝播の社会的影響力に着目し、それに応じた情報伝播挙動の特徴分析を行った。また、ある言語で発生した情報が、他の言語コミュニティに伝播する現象に着目し、こうした言語間の情報伝播挙動の分析を行った。

さらに、メディア間における話題の伝播の時系列を比較し、その類似度や時間差を基に話題の閲覧・探索を行えるインタラクティブな可視化手法を開発し、実装を行った。

４．研究成果

（１）社会的影響力に基づく情報伝播の特徴分析

情報伝播が連鎖することによって引き起こされる情報拡散を情報カスケードと呼ぶ。マイクロブログにおける情報カスケードには早期に対応、あるいは認知するべき情報が含まれるが、一方でアフェリエイトリンクへと誘導するスパムやジョーク、有名人の日常のつぶやき、広告など社会的影響力の少ない情報も多い。注目すべき情報カスケードを検知するための研究としては、将来的に広く拡散する可能性のあるカスケードを検知する手法や、スパムツイートを検知する手法があるが、前者は広く拡散する情報カスケードが必ずしも社会的影響力を持つとは限らない点で、また後者はスパム以外にも社会的影響

力のない情報カスケードが存在するという点で、社会的影響力を持つ情報カスケードを検知する上では不十分である。

本研究では、マイクロブログにおける情報カスケードの中から社会的影響力を持つものを検知するという新しいタスクを設定し、これを教師あり学習に基づく分類器を用いて具体的に解く手法を提案する。社会的影響力を持つ情報カスケードがどのようなものか自明ではないため、我々は「ツイートに含まれる情報が広まることで行動や意思決定に影響を受ける人が存在するか」という観点で社会的影響力を持つ情報カスケードの定義を行い、さらに被験者により社会的影響力を持つとされた情報カスケードを人手で分析することで、マイクロブログ上でどのような社会的影響力を持つ情報カスケードが存在するかを明らかにする。その後、この知見に基づき，社会的影響力を持つ情報カスケードの早期検知を試みる。この問題には、社会的影響力を持つ情報カスケードは全体の約二割程度しか含まれないという難しさと、投稿内容のみからでは分類のための手がかりが十分には得られないという難しさが存在する。そこで本稿ではテキスト特徴量に加えカスケード毎のユーザ分布やカスケードのグラフ構造を特徴量として用いることで社会的影響力を持つ情報カスケードの自動検知を試みる。実験では、分析の際に構築したデータセットを基に SVM を用いて分類器を学習し、分類器の示す F 値によって提案手法の有効性の評価を行った。

本実験での情報カスケードは Twitter API による公式リツイートによって拡散されたツイート(元ツイート)とその公式リツイートの集合とする。分析対象となる情報カスケードは、次で述べるインタラクショングラフに含まれるユーザを観測対象のユーザセットと限定した上で 2013 年 1 月、2 月のツイートそれぞれで 600 回以上リツイートが観測された日本語を含む元ツイートとそのリツイートを抽出することによって作成した。

情報カスケードを抽出する期間以前の 2012 年 1 月から 12 月のユーザ間の投稿のやり取りを元にユーザ間の関係を表す有向グラフ(インタラクショングラフ)G を作成し、情報カスケードの経路を推定する。リツイートとメンションはどちらもユーザ間の情報のやり取りを表しており、このようにして得られるユーザ間のつながりはカスケードの情報伝播の主要な経路となると考えられる。そこで各ユーザをノードとして（過去に）情報が流れる方向と同方向となるよう、リツイートに関してはリツイート元からリツイートしたユーザへ情報が流れるため同方向のエッジを、メンションに関してはメンションを送る際は送り先のユーザの投稿を見て送ったと考えられるためメンションの方向とは逆向きのエッジを追加しインタラクショングラフ G を得る。

こうして得られた情報カスケードに対し、（著者を含まない）3 人の被験者により、「ツイートに書かれた情報を知ったり，その情報を不特定多数に知られたりすることで，直接的あるいは間接的に行動や意思決定に影響を受けるか人がいるか」という観点で社会的影響力の有無を注釈付けした。注釈付けの際には，社会的影響力を持つ情報カスケードの典型例、およびボーダーケースとなる情報カスケードの例を含むアノテーションガイドラインを著者らが作成し、これを参照して実際にリツイートしたユーザが読む/見ると考えられる、元ツイートの本文、画像、元ツイートに含まれる URL のリンク先の情報を基に注釈付けを行ってもらった。これによって得られたラベルにおける三人の被験者間一致度は 0.69 となり、社会的影響力を持つ情報に関してかなりの共通認識が得られていることが確認された。最終的にラベルの不一致は多数決により解消した。その結果を表 1 に示す．表 1 から分かる通り、社会的影響力を持つ情報カスケードは全体のおよそ 20%弱と少ないことが確認された。

表 1　カスケードの社会的影響力の有無

	1 月	2 月
社会的影響力有	188	106
社会的影響力無	942	369
合計	1130	475

本研究では、サポートベクターマシン(SVM)を用いて情報カスケードの社会的影響力の有無を識別する分類器を学習し、社会的影響力を持つ情報カスケードの検知を行う。分類には、どのような内容の情報を、どのような過程で、誰が広めているかが手がかりとなり得る。そこで、これらをカスケードの元ツイートから抽出したテキスト統計量、インタラクショングラフを利用して抽出したカスケードのグラフ特徴量、及び拡散に参加したユーザに関するユーザ特徴量で捉える。

実験では、各カスケードから初期のリツイートを先頭から n 件を取り出し、これを初期カスケードとみなして分類器の学習・評価を行った (n=50，200，400，600)。全ての情報カスケードが社会的影響力を持つと判定する分類器を弱いベースラインとして採用し、提案手法により F 値でこれを上回ることを目標とする。加えて、各特徴量単独で分類を行った場合を強いベースライン（以下、Text、Graph、 User）として、特徴量を組み合わせることの妥当性を確認する。

特徴量を変えて実験を行った際の検知結果を表 2 に示す。参考のため、各被験者と正解ラベル(多数決)との一致度もともに示す。結果として、ユーザ特徴量とテキスト特徴量とグラフ特徴量を同時に用いたとき(All)、カスケードサイズが全ての時点で最も良い

性能を示すことが分かる。また、これらの値はどれも、全てのカスケードが社会的影響力を持つとした場合(Baseline)の F 値 0.45 と比較して顕著な改善であることが分かる。また、カスケードサイズがどの時点でもユーザ特徴量は全ての特徴量を組み合わせた場合と比較しても遜色ない性能を示しているため、提案した特徴量のうちユーザ特徴量が性能において支配的であることがわかる。実用性を考えると、拡散初期の段階で社会的影響力を持つ情報カスケードを認識できることが望ましいが、カスケードサイズを減らし、拡散初期に利用できる特徴量を用いた場合でも、ユーザ特徴量に加えてテキスト特徴量やグラフ特徴量を用いることで、0.576 と Baseline を上回る F 値で社会的影響力の有無を判定できることが示された。

	カスケードサイズ			
特徴量	50	200	400	600
User	0.563	0.594	0.651	0.698
Text	0.552	0.552	0.552	0.552
Graph	0.468	0.521	0.515	0.503
All	**0.576**	**0.614**	**0.661**	**0.709**
Baseline	0.450	0.450	0.450	0.450
被験者 (A)	0.907	0.907	0.907	0.907
被験者 (B)	0.818	0.818	0.818	0.818
被験者 (C)	0.900	0.900	0.900	0.900

表 2　カスケードサイズに伴う精度変化

（2）言語横断的情報カスケードの予測

ソーシャルメディアにおいては、グローバルな多言語によるコミュニケーションが日々行われており、Twitter では、2015 年 12 月時点で 3.2 億のアクティブユーザのうち 79%が米国以外のユーザであり、35 言語がサポートされている。こうしたグローバルなソーシャルメディア上では、ある言語で伝播した情報が、他の言語へカスケードする現象が多くみられ、言語コミュニティ間の情報伝播とみなすことができる。例えば、2014 年に流行した「ALS アイスバケツチャレンジ」は、筋萎縮性側索硬化症の周知と患者への募金を目的として、バケツ一杯の氷水を頭からかぶる画像をソーシャルメディアに共有していく活動で、様々な国に伝播した。こうした多言語横断的な情報伝播を分析し、検知可能とすることは、社会学的な興味のみならず、国際的なマーケティング戦略等を考える際に有用である。

本研究では、情報伝播が成長しかつ言語横断的になる現象を、初期の情報伝播の挙動から予測することを目的とし、Twitter 上の情報伝播におけるユーザの言語分布、成長し言語横断的になる情報伝播の特徴分析を行った。図 1 は、各ユーザが最もよく使う言語(主言語）で書かれたツイートの比率によるユーザ数の累積分布を示したものである。主言語ツイート比率が低いほど他国語を用いる割合が多いことになる。図 1 では、主言語比率が 8 割以下で 2 割以上他国語を使うユーザ（多言語ユーザ）数は約 17%程度であり、83%は 8 割以上主言語しか用いていないユーザ（単一言語ユーザ）であることが示されている。

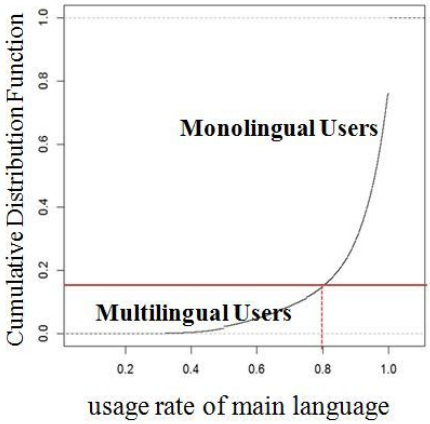

図 1　ユーザの主言語使用比率

本実験においても、（1）の実験と同様に公式リツイートによる情報伝播をカスケードとして取り扱うが、カスケードが言語横断的になるかどうかが本実験の関心事である。あるツイートが k 回リツイートされた時点での観測情報から、最終的な言語横断の度合いを予測することが目的となる。各リツイートにおいてリツイート元とリツイートしたユーザの主言語が異なる場合、そのリツイートが言語横断であるとし、言語横断率は、カスケード内でのリツイート中、言語横断リツイートの比率とする。図 2 は、最初の 10 リツイートを観測した時点での言語横断率を横軸にとり、予測したい最終言語横断率を縦軸にとって各カスケードをプロットしたものである。最終言語横断率は幅広く広がっており、初期の言語横断率のみでの予測は難しいことがわかる。

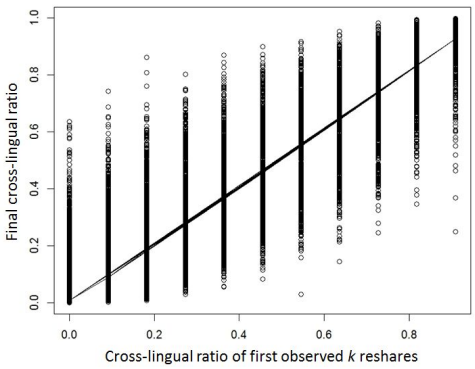

図 2　最終的な言語横断率の分布

言語横断率の予測には、最初のリツイート元のユーザの使用言語の分布（root user)、リツイートしたユーザたちの使用言語の分布(resharer)、ツイートの内容(content)、カスケードに参加したユーザのインタラク

ショングラフの情報(structural)、リツイートの時間間隔(temporal)など多様な手掛かりを用いることを提案し、SVM を用いて、最終的な言語横断比率が一定の閾値を超えるかどうかを予測する問題を解くことで、言語横断的なカスケードを予測するために重要な特徴量を分析した。表 3 は、最終的な言語横断比率が 0.1 を超えるかどうかを予測した場合の精度を示したものである。同時に、各種特徴量を除いた場合の精度変化についても示してある。F 値での評価では、約 0.9 と高い精度で予測ができており、提案した特徴量の多くが有効であることが示された。特に支配的であったのがリツイートしたユーザたちの使用言語分布(resharer)であり、元のユーザよりも、リツイートしたユーザたちの特徴が重要であることが分かる。またツイート内容(content)については、今回多言語を考慮した特徴量を設計できなかったため、除いた場合にむしろ精度が向上しており、効果が無いことが示された。これについては、単語の意味ベクトルを他の言語の意味ベクトルに翻訳する手法等を用いて改善することを考えており、今後の課題として残されている。

表 3　最終言語横断率の予測精度

Features	Accuracy %	Precision	Recall	F-score
Baseline	24.22	0.2422	1	0.3899
SVM(All)	95.12	0.8939	0.9061	0.8999
- root user	94.66	0.8733	0.9118	↓0.8921
- resharer	87.66	0.7728	0.6945	↓0.7315
- content	95.25	0.8992	0.9057	↑0.9024
- structural	95.13	0.8931	0.9074	↑0.9001
- temportal	95.09	0.8946	0.9038	↓0.8991

（3）多メディア間の話題探索のための 3 次元可視化システム

社会現象を分析する際には複数メディア間の話題の広がりを分析することが重要である。多くのメディアでは積極的に映像・画像を用いることで文章だけでは伝えきれない話題・興味の対象を視覚的に伝えており、映像・画像を含めた話題追跡が必要不可欠である。本研究では、複数のメディアから抽出された話題に関する時系列画像群を画像ヒストグラムとして 3 次元空間に可視化し、話題の推移、時系列の差異、メディア間の関係などを視覚的に探索可能にする可視化システムを構築した。

図 3 に構築した可視化システムの全体図を示す。本システムは、我々が文部科学省「多メディア Web 解析基盤の構築及び社会分析ソフトウェアの開発」において構築してきた個別の可視化アプリケーションから、基本構成要素群を抽出及び整理し、様々な応用に適用可能な時系列可視化システムとして統合したものである。ブログ等のソーシャルメディアや、放送映像アーカイブ等の多メディア・リソースから画像情報及びテキスト関連情報が抽出され、画像の時系列については、画像のサムネイルをヒストグラム上に積み上げた ImageHistogram コンポーネントとして表現され、テキストに関する単語頻度等の時系列については、LineChart コンポーネントとして表現され、複数のタイムラインを比較可能な TimeLine コンポーネント上に配置される。ある時間のスナップショットに関する情報は、TimeSlice コンポーネント上に表示される。

こうして可視化された話題の広がりや、メディア間の話題の時間差等によって画像やテキストをフィルタするために、ParallelCoordinateView ダイアログを用いることができ、様々な観点から多メディア間の話題の違いを閲覧することが可能となっている。本統合可視化システムは、ソーシャルメディアや放送映像アーカイブ等を多メディア・リソースとして様々なケーススタディーを行うことで有効性を確認している。

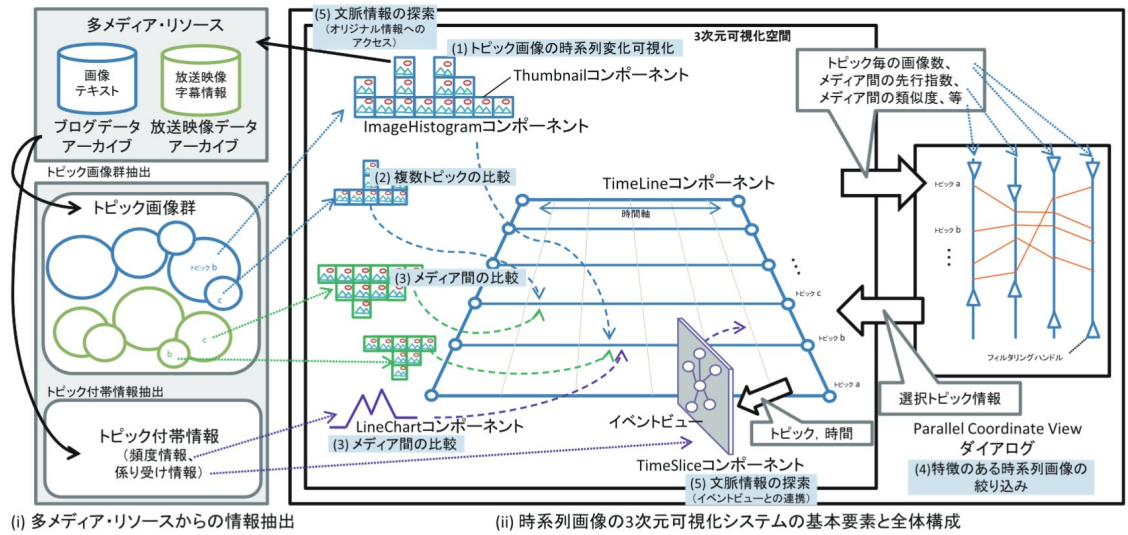

図 3　時系列画像可視化システムの全体構成

５．主な発表論文等

〔雑誌論文〕（計 7 件）

① 川本貴史、豊田正史、吉永直樹、マイクロブログからの社会的影響力を持つ情報カスケードの検知手法、情報処理学会論文誌：データベース(TOD)、査読有、9 巻、2016

② 豊田正史、大規模な階層的グラフのインタラクティブ可視化システム及びその Web メディア分析への応用、可視化情報学会誌、査読無、36(141)、2016、63-67

③ 伊藤正彦、豊田正史、喜連川優、多メディア間の話題探索のための時系列画像 3 次元可視化システム、情報処理学会論文誌:データベース(TOD)、査読有、8 巻、2014、27-44

〔学会発表〕（計 38 件）

① Hongshan Jin、Masashi Toyoda、Analysis of Growing Cross-lingual Cascacdes on Twitter、第 8 回データ工学と情報マネジメントに関するフォーラム(DEIM2016)、2016 年 2 月 29 日～3 月 2 日、ヒルトン福岡シーホーク（福岡県）

② 川本貴史、吉永直樹、豊田正史、マイクロブログにおける社会的影響力を持つ情報カスケードの早期検知に向けて、第 8 回 Web とデータベースに関するフォーラム(WebDB Forum 2015)、2015 年 11 月 24 日～25 日、芝浦工業大学（東京都）

③ Shonosuke Ishiwatari、Nobuhiro Kaji、Naoki Yoshinaga、Masashi Toyoda、Masaru Kitsuregawa、Accurate Cross-lingual Projection between Count-based Word Vectors by Exploiting Translatable Context Pairs、The 19th Conference on Computational Language Learning (CoNLL2015)、2015 年 7 月 30 日～31 日、ベルリン（ドイツ）

④ Masashi Toyoda、Keynote: Multiple Media Analysis and Visualization for Understanding Social Activities、The 4th Temporal Web Analytics Workshop (TempWeb) in conjunction with WWW2014、2014 年 4 月 7 日、ソウル（韓国）

⑤ Masahiko Itoh、Masashi Toyoda、Cai-Zhi Zhu、Shin’ichi Satoh、Masaru Kitsuregawa、Image Flows Visualization for Inter-Media Comparison、IEEE Pacific Visualization (PVis 2014)、2014 年 3 月 4 日～7 日、横浜（日本）

〔図書〕（計 1 件）

① Masashi Toyoda、Masaru Kitsuregawa、Springer、Encyclopedia of Social Network Analysis and Mining、Chapter: Connecting Communities、2014、260-262

〔産業財産権〕

○出願状況（計　0　件）

○取得状況（計　0　件）

〔その他〕

なし

６．研究組織

(1)研究代表者

豊田　正史（TOYODA, Masashi）

東京大学・生産技術研究所・准教授

研究者番号：60447349

(2)連携研究者

伊藤　正彦（ITOH, Masahiko）

東京大学・生産技術研究所・特任准教授

研究者番号：60466422